Sensitivity and specificity can change in opposite directions when new predictive markers are added to risk models

Ben Van Calster[1,2], PhD, Ewout W. Steyerberg[3], PhD,
Ralph B. D'Agostino Sr[2,4], PhD, Michael J. Pencina[2,4], PhD

1. KU Leuven Department of Development and Regeneration, Leuven, Belgium
2. Department of Biostatistics, Boston University, Boston (MA), USA
3. Department of Public Health, Erasmus MC, Rotterdam, the Netherlands
4. Harvard Clinical Research Institute, Boston (MA), USA

Running head: Change in sensitivity and specificity when comparing prediction models

Word count: 3903

Financial support for this study was provided entirely by Research Foundation–Flanders (FWO) (1251609N, 1251612N, G049312N). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. BVC is a postdoctoral fellow of the Research Foundation–Flanders (FWO).

Corresponding author:
Ben Van Calster
KU Leuven, Department of Development and Regeneration
Herestraat 49 box 7003
B-3000 Leuven
Belgium
Tel +32 16 346258
ben.vancalster@med.kuleuven.be

Abstract

When comparing prediction models, it is essential to estimate the magnitude of change in performance rather than rely solely on statistical significance. In this paper we investigate measures that estimate change in classification performance, assuming two-group classification based on a single risk threshold. We study the value of a new biomarker when added to a baseline risk prediction model. First, simulated data are used to investigate the change in sensitivity and specificity (ΔSe and ΔSp). Second, the influence of ΔSe and ΔSp on the Net Reclassification Improvement (NRI; sum of ΔSe and ΔSp) and on decision-analytic measures (Net Benefit or Relative Utility) is studied. We assume normal distributions for the predictors, and assume correctly specified models such that the extended model has a dominating receiver operating characteristic curve relative to the baseline model. Remarkably, we observe that even when a strong marker is added it is possible that either sensitivity (for thresholds below the event rate) or specificity (for thresholds above the event rate) decreases. In these cases decision-analytic measures provide more modest support for improved classification than NRI, even though all measures confirm that adding the marker improved classification accuracy. Our results underscore the necessity of reporting ΔSe and ΔSp separately. When a single summary is desired, decision analytic measures allow for a simple incorporation of the misclassification costs.

Keywords

Biomarkers, decision-analytic measures, net benefit, net reclassification improvement, relative utility, risk assessment, risk factors, sensitivity and specificity

INTRODUCTION

The medical literature is abundant with clinical applications of prediction models to estimate the risk of having (diagnosis) or developing (prognosis) a targeted condition.[1-12] Such models, if successfully validated, can enhance personalized healthcare by supporting individual treatment decisions. With new potentially predictive markers becoming rapidly available from genomics, proteomics, imaging, pathology, blood analysis, and ultrasonography, substantial effort is put into improving risk prediction. This leads to an increasing focus on the evaluation of incremental value of markers added to models[13,14] and in the comparison of competing prediction models.

When assessing the incremental value of an added marker, recent literature advises only one formal test of hypothesis to assess the statistical significance of the marker in a multivariable regression model.[15-17] After the significance has been established, measures quantifying the predictive performance of risk models are important to estimate the magnitude of performance improvement. These can be based on the model-based risks to provide a "global" assessment[18], or on risk categories obtained by using thresholds to classify patients into two or more groups[19,20]. Using two groups (low vs. high risk) based on a single threshold, the rule might be used to indicate whether or not patients should receive treatment. Using two thresholds creates an intermediate risk group which may be used to select patients for less invasive treatment or additional testing.

The most common "global" measure is the c-statistic, known as the area under the receiver operating characteristic (ROC) curve (AUC) for dichotomous outcomes. The

AUC difference between models, ΔAUC, has been criticized as being insensitive and hard to interpret, motivating novel measures such as Integrated Discrimination Improvement and continuous Net Reclassification Improvement.[18,21-23] To evaluate improved classification, a widely used measure is the Net Reclassification Improvement (NRI).[21,24] NRI is the sum of the net percentage of events reclassified to a higher risk group and the net percentage of non-events reclassified to a lower risk group. For classification in two groups, NRI equals the sum of the differences in sensitivity and specificity or the difference in the Youden index.[21]

Simultaneously, decision-analytic measures accounting for different misclassification costs of events and non-events are gaining attention.[25-28] The main measures are the increase in Net Benefit (ΔNB), increase in Relative Utility (ΔRU), and weighted NRI.[22,29,30] These three measures are transformations of one another and thus always favor the same model.[31] Similar to the two-group NRI, they can be seen as functions of changes in sensitivity and specificity. According to principles of decision analysis, the risk threshold used for classification defines the relative misclassification costs[32]: the odds at threshold equal the ratio of the harm of a false positive to the benefit of a true positive. For example, a threshold of 20% implies 1:4 odds and thus classifying one true positive is worth misclassifying four false positives. This ratio is incorporated into the decision-analytic measures. On the other hand, the NRI assumes that the cost ratio is a function of the event rate and equals the odds of non-events. It is hence theoretically possible that the decision-analytic measures do not favor the same model as the NRI.[31,33]

When one model has a higher sensitivity and specificity than the other, all measures will favor this model. However, NRI and decision-analytic measures may differ in their value

depending on the adopted risk threshold and the specific changes in sensitivity and specificity (ΔSe and ΔSp). When sensitivity is increased but specificity decreased, the difference between the two-group NRI and decision-analytic measures may become more substantial. Decreasing sensitivity or specificity appears counterintuitive in the case of a "useful" marker being added to a baseline model, yet recently such a counterintuitive result was reported.[34]

In this paper we extend these results by investigating ΔSe and ΔSp under various scenarios in which two models are compared, and by studying the differential influence of ΔSe and ΔSp on the NRI and decision-analytic alternatives. We focus on ΔSe and ΔSp as a function of a single risk threshold, and the agreement between the two-group NRI and decision-analytic measures. We mainly rely on simulated data involving correctly specified models, but illustrate our findings on prediction of coronary heart disease in the Framingham data[21].

METHODS

Simulations

We address three general scenarios: adding a continuous (I) or binary (II) marker to a baseline model, and the comparison of a baseline model to a non-nested competing model (III). The marker and the set of predictors in a model are each represented by one variable. In scenarios I and II we compare a baseline logistic regression model based on linear predictor X_1 with an extended model where the marker X_2 is added. In scenario III

we compare two models based on two linear predictors X_1 and X_2. For scenarios I and III we assume X_1 and X_2 to be normally distributed among events and non-events (with standard deviation of 1), and, without loss of generality, set the mean among non-events at 0. Given that under normality means and AUCs are mathematically related, the mean of X_1 and X_2 among events is chosen to achieve a specific AUC and ΔAUC.[35] Scenario II differs from scenario I in that X_2 is a binary marker with prevalence among non-events and events chosen to achieve a specific ΔAUC. For each scenario, we define a main setting as well as a set of variations on that. We assume that the models' estimated risks are well calibrated and that models are correctly specified. This implies that the extended or competitor model has a dominant ROC curve relative to the baseline model. With calibrated risks we mean that the estimated risks correspond to observed proportions: among women with an estimated risk of event of 0.3, 30% are expected to have the event. In what follows, the first scenario is described in detail. For the other two scenarios, comparable methods are being used which are presented in Web Appendix 1. Specific differences are the issues of marker prevalence in scenario II and of correlation between models in scenario III.

For the main setting of scenario I, means for X_1 and X_2 are derived to reflect a baseline model with an AUC of 0.7 to which a marker is added that increases AUC by 0.05. The correlation between X_1 and X_2 is set to 0, and the event rate to 10%. The following basic variations are considered: (a) the baseline model has a very strong discrimination, i.e. AUC = 0.8, while keeping ΔAUC at 0.05; (b) the added marker increases the AUC by 0.01; (c) the added marker increases the AUC by 0.1; (d) event rate is 1%; (e) event rate is 30%; (f) event rate is 70%. Due space constraints we address each variation separately. We did address many possible combinations of these variations but did not obtain

meaningfully different results. Table 1 provides an overview of the main setting and the variations, including AUCs and ΔAUC.

To obtain stable values for model performance on the population level, we simulated a dataset containing at least 5 million patients for each setting. We fitted the two models, computed classification performance and improvement for thresholds varying from 1% to 99% with increments of 1%. Measures for classification improvement included ΔSe and ΔSp, two-group NRI, ΔRU, and ΔNB/event rate. NB corrects the proportion of true positives for the proportion of false positives weighted according to the adopted misclassification costs, i.e. odds of the risk threshold. NB can be written as follows:

$$\mathrm{NB} = \frac{\#\mathrm{TP} - \#\mathrm{FP} * \mathrm{odds}(\mathrm{threshold})}{N}, \quad (1)$$

and ΔNB as:

$$\Delta\mathrm{NB} = \frac{\Delta\mathrm{TP} - \Delta\mathrm{FP} * \mathrm{odds}(\mathrm{threshold})}{N}, \quad (2)$$

NB is typically compared to the net benefit of two default strategies: 'treat all' (i.e. consider everyone high risk) or 'treat none' (consider everyone low risk). Finally, RU is the proportion of the maximum attainable net benefit that the model captures over the best default strategy at the adopted threshold, $\mathrm{NB}_{default}$. ΔRU can be written as:

$$\Delta\mathrm{RU} = \frac{\Delta\mathrm{NB}}{\text{event rate} - \mathrm{NB}_{default}}. \quad (3)$$

Note that the best default strategy is 'treat all' for thresholds below event rate, and 'treat none' for thresholds above event rate.

We report ΔRU and ΔNB/event rate because these can be expressed as weighted sums of ΔSp and ΔSe and can therefore be compared with NRI as the standard sum of ΔSp and ΔSe. ΔNB/event rate is the sum of ΔSe and 'ΔSp weighted for misclassification costs and event rate' and can thus be seen as an increase in net sensitivity: equivalent to the increase in sensitivity at unchanged specificity .[31] For thresholds above event rate, ΔRU is identical to ΔNB/event rate. Else, ΔRU is the sum of ΔSp and 'ΔSe weighted for misclassification costs and event rate', thus an increase in net specificity.[31] ΔSe is given a higher weight than ΔSp for thresholds below event rate, and lower weight for thresholds above event rate. Whether we frame the decision-analytic performance in terms of sensitivity or specificity will have an impact on the resulting value. See Web Appendix 2 for formulas.

Case study: risk prediction for coronary heart disease

A sample of 3264 Framingham Heart Study men and women between 30 and 74 years of age without evidence of cardiovascular disease were included in this analysis. They attended their baseline examination between 1987 and 1992 and their risk factors were collected. The participants were followed for 10 years for development of coronary heart disease (CHD), which included myocardial infarction, angina pectoris, coronary insufficiency or CHD death. A Cox model was fitted to obtain the 10-year risk of CHD,

using age, sex, diabetes, smoking, and systolic blood pressure as predictors. The incremental value of HDL cholesterol is then assessed by adding it to the model[21].

RESULTS

Added continuous marker (scenario I)

Interplay of changes in sensitivity and specificity. The main setting for the continuous maker scenario yielded a completely dominant ROC curve for the extended model (Figure 1a). The risk distribution of the extended model had a lower peak and a heavier tail showing more cases with increased risks (Figure 1b). However, hardly any risks were above 0.5, and hence we used 0.5 as the maximum threshold in the remaining graphs. Despite the dominant ROC curve, sensitivity and specificity did not increase at every classification threshold (Figures 1c-d). For a subset of thresholds between 0 and the event rate, sensitivity decreased when the marker was added. Analogously, specificity decreased with thresholds in the middle between event rate and 100%. The decrease in performance for events was accompanied by a strong increase in the performance for non-events, and vice versa. For thresholds near the event rate, ΔSe and ΔSp were both positive. Variations of the main setting, involving changes in the event rate or the strength of the baseline model or added marker, yielded analogous results that were merely strengthened or weakened depending on the specific variation (Figure 2). When varying event rate (Figure 2d-f), we observed similar decreases (up to -2.3%) and increases (up to 12%) in sensitivity and specificity. The single difference is that results are horizontally shifted to accommodate the different event rates.

Impact on summary measures. The interplay of ΔSe and ΔSp created a bimodal distribution for the NRI. In all settings the NRI and decision-analytic measures were non-negative for all thresholds, and thus acknowledged increased performance when adding the new marker despite decreasing ΔSe or ΔSp for some thresholds.

When ΔSe or ΔSp was negative, decision-analytic measures gave a numerically more modest view of the improved classification. The extent to which the decision-analytic measures were more modest depended on whether the improved classification was expressed as an increase in net sensitivity or net specificity. If ΔSe was negative, the increase was more modest when expressed in terms of net sensitivity. This can be explained by the fact that for these thresholds ΔSe was given more relative weight compared to ΔSp. Likewise, if ΔSp was negative decision-analytic measures were more modest when expressed in terms of net specificity. When ΔSe or ΔSp were both positive, the decision-analytic value expressed as the increase in net sensitivity was lower than the NRI when the threshold is lower than the event rate but higher for thresholds above the event rate. The opposite was true when the decision-analytic measure was expressed as the net increase in specificity. By definition, the NRI and the decision-analytic measures had similar values for thresholds at or near the event rate.

Added binary marker (scenario II)

The main setting now involves a binary marker X_2 with a prevalence of 15% in non-events and 41% in events to achieve a ΔAUC of 0.05. For this setting and its variations, the results were very similar to the previous scenario (Web Figures 1-2). We found that

ΔSe or ΔSp were often negative despite a dominant ROC curve for the extended model, with similar effects of the interplay of ΔSe and ΔSp on the NRI versus decision-analytic measures. Regarding marker prevalence, the lower it is the larger the range of thresholds for which sensitivity decreases. With prevalence of 1% in non-events and 19% in events, sensitivity decreased even for the threshold equal to the event rate.

Comparison of non-nested models (scenario III)

Results for scenario III were entirely comparable to those of the other scenarios (Web Figures 3-4).

Case study on CHD risk prediction

The 10-year CHD event rate was 5.6%. Adding HDL cholesterol to the model resulted in an adjusted hazard ratio of 0.65 per standard deviation (95% confidence interval: 0.53, 0.80), suggesting a clear protective association with the risk of CHD. The baseline model had an AUC of 0.762, which increased to 0.774 when HDL cholesterol was added (ΔAUC = 0.012). Classification and difference in classification results for a few possible thresholds are consistent with the simulation results although some confidence intervals span zero (Table 2). When the common threshold of 20% was used, adding HDL increased sensitivity but mildly decreased specificity . The odds of this threshold are 1:4, meaning that treating one individual that would develop CHD within 10 years is worth the unnecessary treatment of 4 individuals who would not develop CHD. For these relative misclassification costs, the improvement in classification performance was equivalent to a net increase in sensitivity of 5.3% at unchanged specificity

(ΔNB/event rate = 0.053). Or, when expressed as the net increase in specificity (Web Appendix 2), the improvement was equivalent to an increase in specificity of 1.3% at unchanged sensitivity. NRI was positive as well and thus also suggested improved classification.

At a threshold of 6%, which has also been advocated, one true positive is worth 16 false positives. This threshold was associated with improved sensitivity and mildly improved specificity. This was expected since 6% is close to the event rate of 5.6%. At a threshold of 4%, for which one true positive is worth 24 false positives, sensitivity worsened whereas specificity improved.

DISCUSSION

In this paper, we investigated the difference in sensitivity and specificity under various scenarios that involve a comparison of predictions models. In addition, we elucidate if and how NRI and decision-analytic measures may lead to different conclusions. In contrast with intuition, we found that a decrease in sensitivity or specificity often occurs at specific thresholds, even when a marker with strong incremental value is added to a model or when a competing model has a clearly superior discrimination. At the same time the decrease in sensitivity (specificity) will be more than compensated by an increase in specificity (sensitivity). This finding was seen for a wide range of settings, hence consolidating a similar coincidental finding in a recent study.[34] Thus it is important to understand that a possible decrease in sensitivity or specificity does not necessarily compromise the marker's incremental value. As a result, it is more

informative to report the differences in sensitivity and specificity separately rather than their sum in the form of the NRI.[36] Although this recommendation was already made in the paper that proposed the NRI[21], further emphasis is required since many current reports focus on the combined NRI rather than its components.

We submit that the robustness of the observed phenomenon was unexpected to us, and it appears to have more to do with mathematics than intuition. However, it is logical that specificity improves more strongly than sensitivity for low risk thresholds and vice versa for high thresholds. The use of a low risk threshold suggests a relative preference for true positives over true negatives, and will lead to a relatively high sensitivity for the baseline model. This implies that adding a marker leaves more room to improve specificity than sensitivity. The phenomenon of decreasing sensitivity or specificity becomes less apparent when the discrimination performance of the baseline model is better but also when the discrimination performance of the added marker or competitor model is better (panels A and C in Figure 1 and Web Figures 2 and 4): eventually, higher discrimination will by definition lead to perfect sensitivity and specificity.

At the population level, the NRI and decision-analytic alternatives are likely to be different in magnitude mainly when ΔSe and ΔSp are different in sign, which usually happens when the threshold is in the middle between event rate and either 0 or 1. If a single summary measure is desired in the two-group setting, we recommend presenting such measures alongside differences in sensitivity and specificity, given that decision-analytic measures allow explicit incorporation of the misclassification costs. These performance measures are essential to estimate the magnitude of change in classification performance, whereas formal hypothesis testing of the incremental value

of an added marker should focus on the marker's coefficient in a multivariable regression model.[15-17]

The decision-analytic measures can be framed as a difference in net sensitivity or as a difference in net specificity[31], and we believe that it is useful to express these measures from both viewpoints. In this work we used NB/event rate rather than NB itself in order to better compare it with NRI and RU. While NB/event rate can be seen as a net sensitivity, NB expresses the result on an absolute scale as the net proportion of true positives. The absolute expression can be seen as clinically more relevant as it is not conditional on the event rate but rather expresses the result at the level of the patients themselves. Even then, NB can be reworked to give the net proportion of true negatives. On the contrary, RU calculations depend on whether the threshold is below or above the event rate. In the former case the default strategy is 'treat all' and RU is expressed in the form of a net specificity, in the latter case it is the other way around. This makes sense, because 'treat all' has 100% sensitivity and 0% specificity such that comparing a model with this default strategy is logically done on the specificity scale.

We acknowledge that performing simulation studies under the assumption of normality of predictors and based on one specific model type (logistic regression) may not always generalize to situations where this assumption is violated. We do believe, however, that our results are generalizable to many settings. First, the assumption of normality for a model and for a single added continuous marker is reasonable. Linear predictors for prediction models quite often approach a normal distribution for events and non-events. Also, non-normally distributed markers can very often be made normally distributed through a simple transformation. Note that the assumption of normality has frequently

being used in similar work.[15,18,37] Second, one of the addressed scenarios involved the addition of a binary marker to a prediction model. Results for this scenario were similar to those for other scenarios. Third, under normality with equal standard deviations among events and non-events and with equal correlation between baseline model and marker for events and non-events, the logistic regression is the correct mathematical formulation of the link between predictors and outcome.[37] Fourth, we performed a set of simulations under different conditions (1 - added marker is lognormally distributed among events and non-events; 2 - baseline model consists of four binary markers and added marker is binary as well) and using different model types (probit regression, Poisson regression, and support vector machines), and obtained similar results (Web Appendix 3).

A drawback of this work is that we focused on correctly specified models. Misspecification can refer to lack of calibrated risks, inappropriate modeling of predictors with strongly nonlinear effects, or omission of an important interaction effect. Full investigation of misspecification is a subject on its own that would dilute the main message of the current work. We did, however, carry out simulations under a limited set of misspecification settings. More specifically we investigated situations where the model with an added continuous marker overfits risks, misses a quadratic effect of the marker, or missed an interaction between the marker and information in the baseline model (Web Figure 5). These settings generally showed a similar pattern for ΔSe and ΔSp, with one of them becoming negative for some risk thresholds. ΔNB was always lower under misspecification compared to correct specification, with even negative values observed for some risk thresholds in all three settings. NRI was higher when the model with the added marker overfitted the risk as compared to a well calibrated model,

especially as the risk threshold moved away from event rate. This undesirable result is in line with previous reports.[38]

This work focused on results on the population level by considering an enormous sample size, and suggested that NRI and decision-analytic measures are consistent in sign for correctly specified models. These results give a reference of what may be obtained in studies on the incremental value of novel markers or on the performance of non-nested models. For finite sample sizes (e.g. the presented case study), however, results can be uncertain and inconsistency in sign may nevertheless occur.[31,33] To demonstrate this we simulated 1000 datasets containing 1000 patients for the main setting of scenario 1, and computed how often the NRI and decision-analytic measures were inconsistent in sign (Figure 3). The proportion of inconsistencies was near 0 when the threshold equaled the true event rate, although due to sampling variability of the event rate the proportion was not exactly 0. When the threshold was further from the true event rate, the proportion increased. When the threshold approaches 0 or 1, all measures of improved classification approach 0 by definition and hence the proportion of inconsistencies does too.

Often, more than two thresholds are defined to delineate three or more risk groups to allow several options for managing patients.[24] Evaluation of improved classification is more complex in such cases and, although possible, the extensions of decision-analytic measures to these situations are not straightforward. As a result, the applicability of decision-analytic measures in addition to the NRI components, ΔSe and ΔSp, is clear in the two-group setting, but it needs further investigation in the multiple-group setting.

CONCLUSION

When comparing prediction models, it is essential to estimate the magnitude of change in performance rather than rely on statistical significance. In contrast with what could be expected intuitively, sensitivity or specificity may decrease even when the ROC curve of one model uniformly dominates the ROC curve of the other model. The NRI and decision-analytic measures will agree in sign in reasonable scenarios. But the latter measures will generally give a numerically more modest impression of the improved classification performance of the added marker or competing model. When estimating the difference in the two-group classification performance of two risk models, we recommend reporting the differences in sensitivity and specificity separately. When a combined measure is desired, decision-analytic measures such as NB and RU have the advantage of explicit incorporation of costs.

References

1. Van Holsbeke C, Van Calster B, Bourne T, Ajossa S, Testa AC, Guerriero S, et al. External validation of diagnostic models to estimate the risk of malignancy in adnexal masses. Clin Cancer Res. 2012;18:815-25.
2. Meigs JB, Schrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. New Engl J Med. 2008;359:2208-19.
3. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med. 2012;9:1-12.
4. Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB Sr, et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. Lancet. 2009;373:739-45.
5. D'Agostino RB Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008;117:743-53.
6. Roobol MJ, Schröder FH, Hugosson J, Jones JS, Kattan MW, Klein EA, et al. Importance of prostate volume in the European Randomised Study of Screening for Prostate Cancer (ERSPC) risk calculators: results from the prostate biopsy collaborative group. World J Urol. 2012;30:149-55.
7. Genders TSS, Steyerberg EW, Hunink MGM, Nieman K, Galema TW, Mollet NR, et al. Prediction model to estimate presence of coronary artery disease: retrospective pooled analysis of existing cohorts. BMJ. 2012;344:e3485.

8. Eagle KA, Lim MJ, Dabbous OH, Pieper KS, Goldberg RJ, Van de Werf F, et al. A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry. JAMA. 2004;291:2727-33.

9. Van Calster B, Abdallah Y, Guha S, Kirk E, Van Hoorde K, Condous G, et al. Rationalizing the management of pregnancies of unknown location: temporal and external validation of a risk prediction model on 1962 pregnancies. Hum Reprod. 2013;28:609-16.

10. Van Belle V, Van Calster B, Brouckaert O, Vanden Bempt I, Pintens S, Harvey V, et al. Qualitative assessment of the progesterone receptor and HER2 improves the Nottingham Prognostic Index up to 5 years after breast cancer diagnosis. J Clin Oncol. 2010;28:4129-34.

11. Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D, et al. A risk score for predicting near-term incidence of hypertension: the Framingham Heart Study. Ann Intern Med. 2008;148:102-10.

12. Kastrinos F, Steyerberg EW, Balmaña J, Mercado R, Gallinger S, Haile R, et al. Comparison of the clinical prediction model PREMM1,2,6 and molecular testing for the systematic identification of Lynch syndrome in colorectal cancer. Gut. 2013;62:272-9.

13. Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MS, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation. 2009;119:2408-16.

14. Tzoulaki I, Liberopoulos G, Ioanindis JPA. Assessment of claims of improved prediction beyond the Framingham risk score. JAMA. 2009;302:2345-52.

15. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. BMC Med Res Methodol. 2011;11:13.

16. Pepe MS, Kerr KF, Longton G, Wang Z. Testing for improvement in prediction model performance. Stat Med. 2013;32:1467-82.

17. Demler OV, Pencina MJ, D'Agostino RB Sr. Misuse of DeLong test to compare AUCs for nested models. Stat Med. 2012;31:2577-87.

18. Pencina MJ, D'Agostino RB Sr, Pencina KM, Janssens ACJW, Greenland P. Interpreting incremental value of markers added to risk prediction models. Am J Epidemiol. 2012;176:473-81.

19. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gönen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128-38.

20. Steyerberg EW, Van Calster B, Pencina MJ. Performance Measures for Prediction Models and Markers: Evaluation of Predictions and Classifications. Rev Esp Cardiol. 2011;64:788-94.

21. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27:157-72.

22. Pencina MJ, D'Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011;30:11-21.

23. Kerr KF, Bansal A, Pepe MS. Further insight into the incremental value of new markers: the interpretation of performance measures and the importance of clinical context. Am J Epidemiol. 2012;176:482-7.

24. Mühlenbruch K, Heraclides A, Steyerberg EW, Joost HG, Boeing H, Schulze MB. Assessing improvement in disease prediction using net reclassification improvement: impact of risk cut-offs and number of risk categories. Eur J Epidemiol. 2013;28:25-33.
25. Vickers AJ. Prediction models: revolutionary in principle, but do they do more good than harm? [editorial]. J Clin Oncol. 2011;29:2951-2.
26. Baker SG. Putting risk prediction in perspective: relative utility curves. J Natl Cancer Inst. 2009;101:1538-42.
27. Localio AR, Goodman S. Beyond the Usual Prediction Accuracy Metrics: Reporting Results for Clinical Decision Making [editorial]. Ann Intern Med. 2012;157:294-5.
28. Khalili D, Hadaegh F, Soori H, Steyerberg EW, Bozorgmanesh M, Azizi F. Clinical usefulness of the Framingham cardiovascular risk profile beyond its statistical performance: the Tehran Lipid and Glucose Study. Am J Epidemiol. 2012;176:177-86.
29. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565-74.
30. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. J R Stat Soc A. 2009;172:729-48.
31. Van Calster B, Vickers AJ, Pencina MJ, Baker SG, Timmerman D, Steyerberg EW. Evaluation of Markers and Risk Prediction Models: Overview of Relationships between NRI and Decision-Analytic Measures. Med Decis Making. 2013;33:490-501.
32. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. New Engl J Med. 1975;293:229-34.

33. Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. Eur J Clin Invest. 2012;42:216-28.

34. Mihaescu R, van Zitteren M, van Hoek M, Sijbrands EJ, Uitterlinden AG, Witteman JC, et al. Improvement of risk prediction by genomic profiling: reclassification measures versus the area under the receiver operating characteristic curve. Am J Epidemiol. 2010;172:353-61.

35. Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. J Am Stat Assoc. 1993;88:1350-5.

36. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. Am J Epidemiol. 2011;173:1327-35.

37. Bansal A, Pepe MS. When does combining markers improve classification performance and what are implications for practice? Stat Med. 2013;32:1877-92.

38. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. [published online ahead of print April 2, 2013]. Stat Med.

39. Zou G. A modified Poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004;159:702-6.

40. Van Gestel T, Suykens JAK, Lanckriet G, Lambrechts A, De Moor B, Vandewalle J. Bayesian framework for least squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis. Neural Comput. 2002;14:1115-47.

41. Van Calster B, Timmerman D, Lu C, Suykens JAK, Valentin L, Van Holsbeke C, et al. Preoperative diagnosis of ovarian tumors using Bayesian kernel-based methods. Ultrasound Obstet Gynecol. 2007;29:496-504.

42. Van Calster B, Condous G, Kirk E, Bourne T, Timmerman D, Van Huffel S. An application of methods for the probabilistic three-class classification of pregnancies of unknown location. Artif Intell Med. 2009;46:139-54.

Table 1. Settings for Assessing the Added Value of a Continuous Marker.

Event rate	Mean of X_1	Mean of X_2	AUC Baseline model	AUC Extended model	ΔAUC
Main setting					
10%	0.742	0.6	0.70	0.75	0.05
Variations					
Stronger baseline model					
10%	1.19	0.86	0.80	0.85	0.05
Weaker or stronger added marker					
10%	0.742	0.25	0.70	0.71	0.01
10%	0.742	0.931	0.70	0.80	0.1
Higher or lower event rate					
1%	0.742	0.6	0.70	0.75	0.05
30%	0.742	0.6	0.70	0.75	0.05
70%	0.742	0.6	0.70	0.75	0.05

AUC, area under the receiver operating characteristic curve; ΔAUC, difference in area under the receiver operating characteristic curve

Table 2. Added Value of HDL Cholesterol in the Prediction of Coronary Heart Disease.

	Risk threshold		
	4%	**6%**	**20%**
Baseline model			
Sensitivity	87%	70%	13%
Specificity	51%	68%	97%
Classification improvement when adding HDL (95% CI)			
ΔSens	-2.7 (-6.3; 0.6)	6.0 (1.2; 11.1)	6.0 (2.3; 10.9)
ΔSpec	2.1 (1.0; 3.3)	0.2 (-0.9; 1.4)	-0.2 (-0.7; 0.3)
NRI	-0.006 (-0.045; 0.030)	0.062 (0.013; 0.113)	0.058 (0.020; 0.108)
ΔRU	-0.018 (-0.071; 0.032)	0.063 (0.013; 0.113)	0.053 (0.008; 0.106)
ΔNB/event rate	-0.012 (-0.050; 0.023)	0.063 (0.013; 0.113)	0.053 (0.008; 0.106)

CI, confidence interval; ΔSens, difference in sensitivity; ΔSens, difference in specificity; NRI, net reclassification improvement; ΔNB, difference in net benefit; ΔRU, difference in relative utility.

Figure Legends

Figure 1. Adding a continuous marker to a baseline prediction model: ROC curves (A), risk distributions (B), sensitivity and specificity by risk threshold (C), measures to assess improvement in classification (D). The event rate is 10%, the AUC of the baseline model is 0.70 and of the model with the marker is 0.75.

Figure 2. Adding a continuous marker to a baseline prediction model: assessment of improved classification for variations of the main setting. (A) baseline model has strong discrimination, (B) added marker increases AUC by 0.01, (C) added marker increases AUC by 0.1, (D) event rate 1%, (E) event rate 30% (F) event rate 70%.

Figure 3. Adding a continuous marker to a baseline prediction model: finite sample results showing the proportion of 1000 bootstrap samples where NRI and ΔNB differed in sign as a function of risk threshold.

Web Figure 1. Adding a binary marker to a baseline prediction model: ROC curves (A), risk distributions (B), sensitivity and specificity by risk threshlod (C), measures to assess improvement in classification (D). The event rate is 10%, the AUC of the baseline model is 0.70 and of the model with the marker is 0.75 (marker prevalence is 15% among non-events and 41% among events).

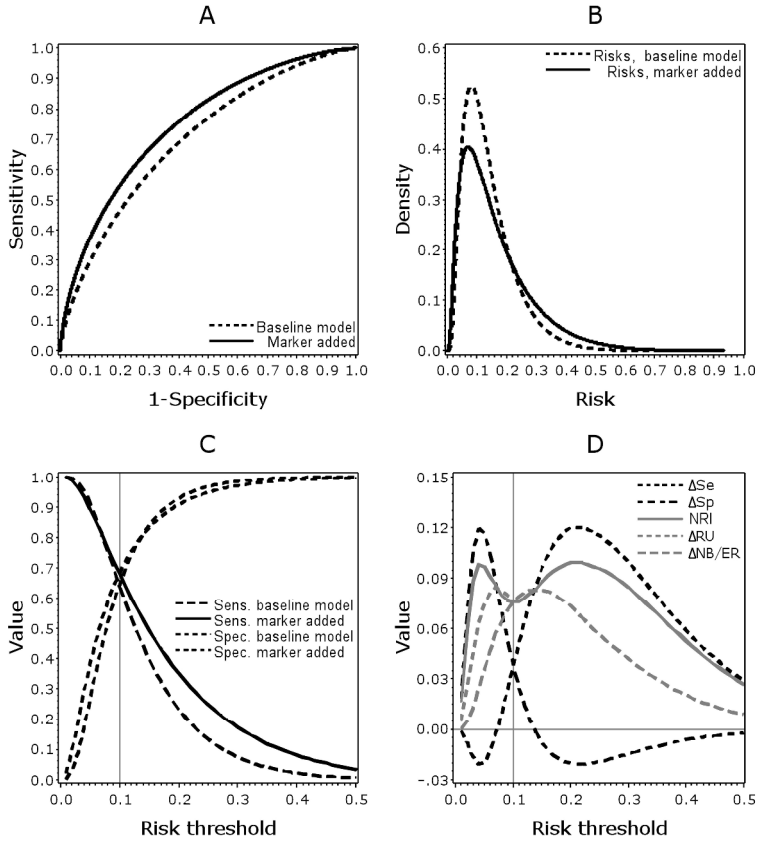
Web Figure 2. Adding a binary marker to a baseline prediction model: assessment of improved classification for variations of the main setting. (A) baseline model has strong discrimination, (B) added marker increases AUC by 0.01, (C) added marker increases AUC by 0.1, (D) event rate 1%, (E) event rate 30%, (F) event rate 70%, (G) prevalence of

binary marker is 50% for non-events while maintaining ΔAUC of 0.5, (H) prevalence of binary marker is 1% for non-events while maintaining ΔAUC of 0.5.

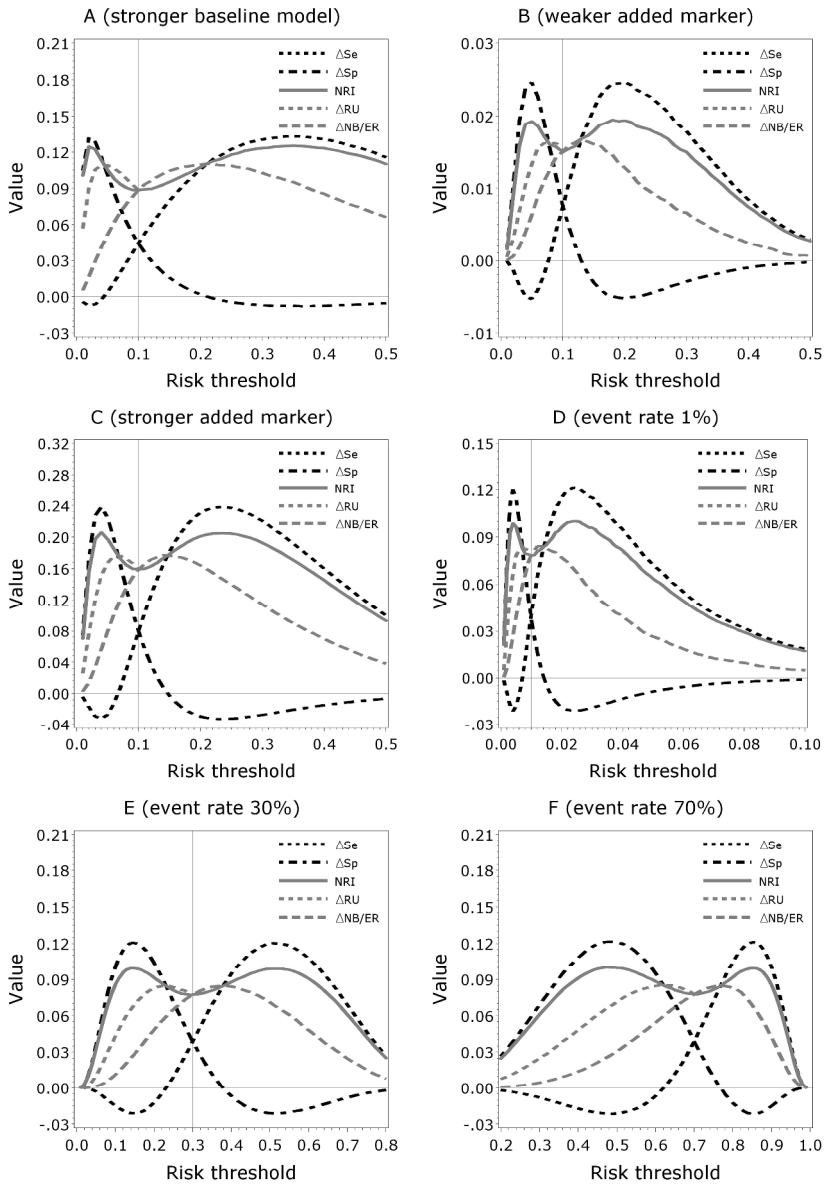
Web Figure 3. Comparing two non-nested models: ROC curves (A), risk distributions (B), sensitivity and specificity by risk threshold (C), measures to assess improvement in classification (D). The event rate is 10%, the AUC of the baseline model is 0.70 and of the competitor model is 0.75, and the correlation between the linear predictors of the two models is 0.5.

Web Figure 4. Comparing two competing models: assessment of improved classification for variations of the main setting. (A) baseline model has strong discrimination, (B) ΔAUC is 0.01, (C) ΔAUC is 0.1, (D) event rate 1%, (E) event rate 30% (F) event rate 70%, (G) competing models are strongly correlated (correlation 0.75).

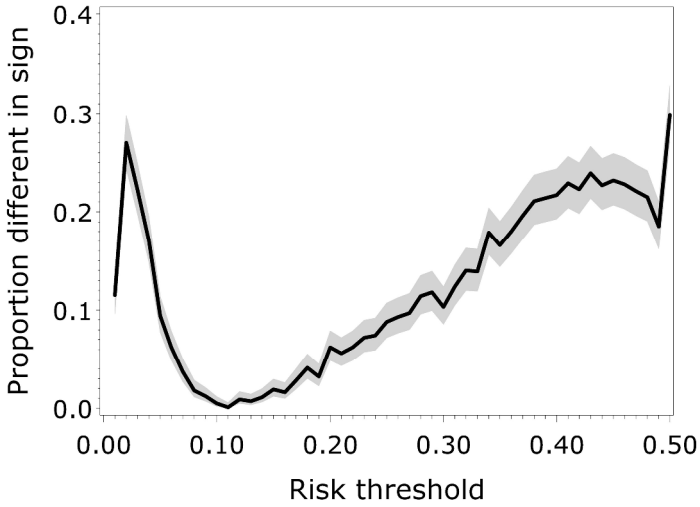
Web Figure 5. Adding a continuous marker to a baseline prediction model when the model with marker is misspecified: assessment of improved classification. (A) Extended model is overfit, (B) Missed interaction effect between marker and baseline model as well as missed quadratic effects for marker and baseline model (effects introduced by assuming correlation of marker with baseline model in events), (C) Missed quadratic effect of marker (effect introduced by increasing SD of added marker for non-events to 2).



284x409mm (300 x 300 DPI)

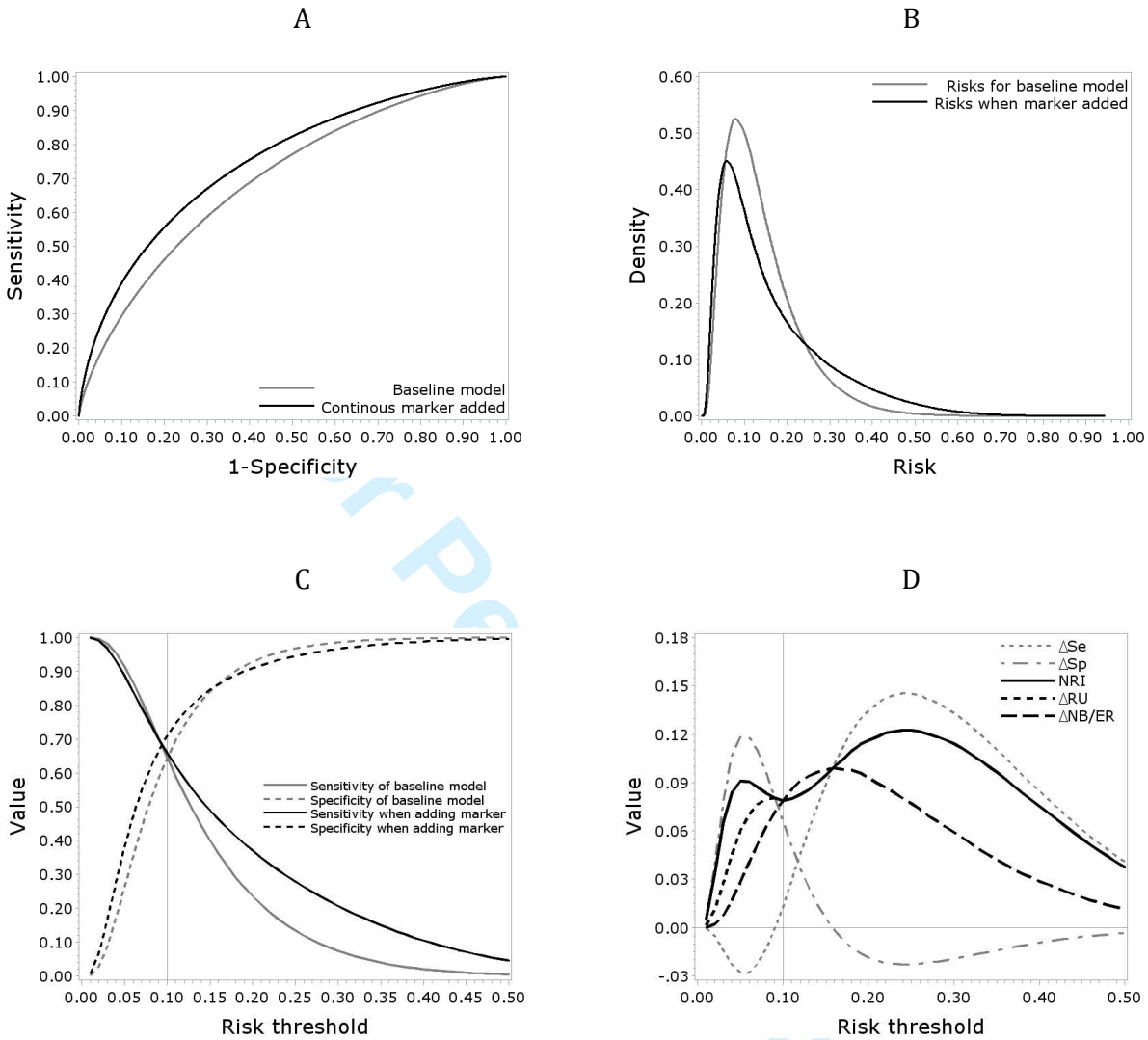


284x409mm (300 x 300 DPI)



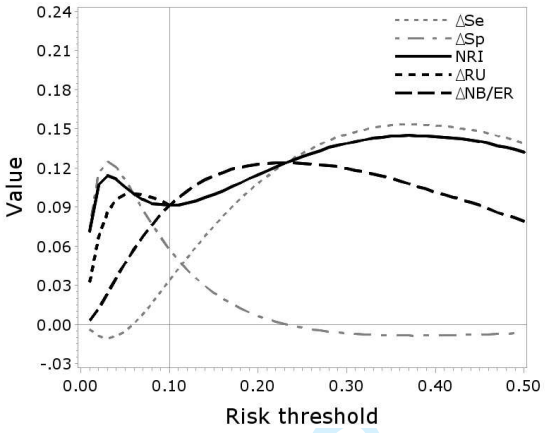
284x409mm (300 x 300 DPI)

Web Figure 1.

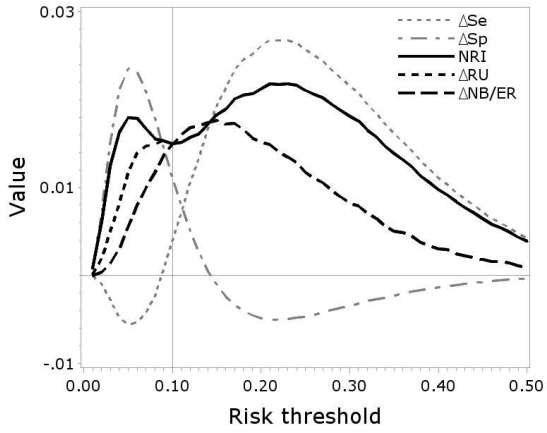

A
Sensitivity
1-Specificity
Baseline model
Continous marker added
B
Density
Risk
Risks for baseline model
Risks when marker added
C
Value
Risk threshold
Sensitivity of baseline model
Specificity of baseline model
Sensitivity when adding marker
Specificity when adding marker
D
Value
Risk threshold
ΔSe
ΔSp
NRI
ΔRU
ΔNB/ER

Web Figure 2.

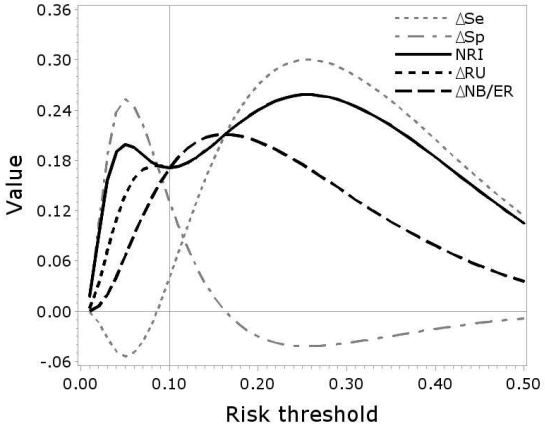
A (stronger baseline model)



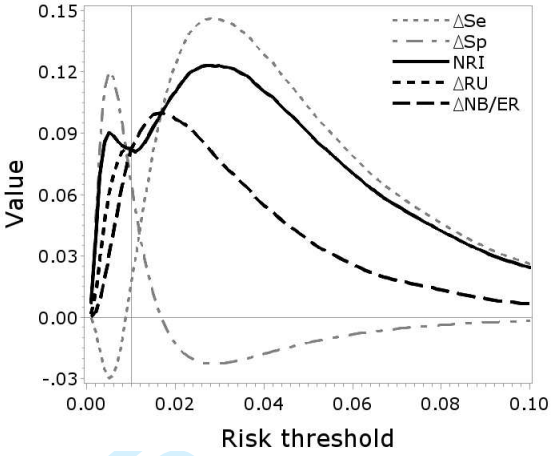
B (weaker added marker)



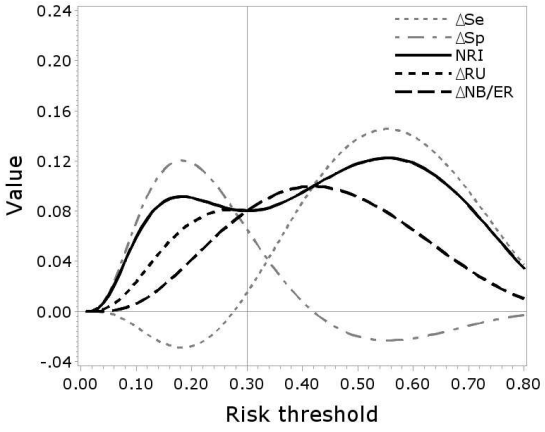
C (stronger added marker)



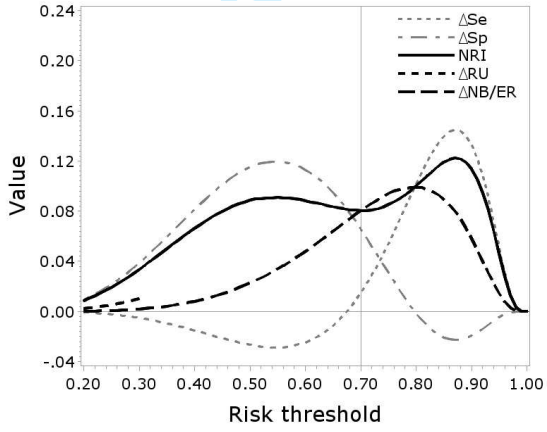
D (event rate 1%)



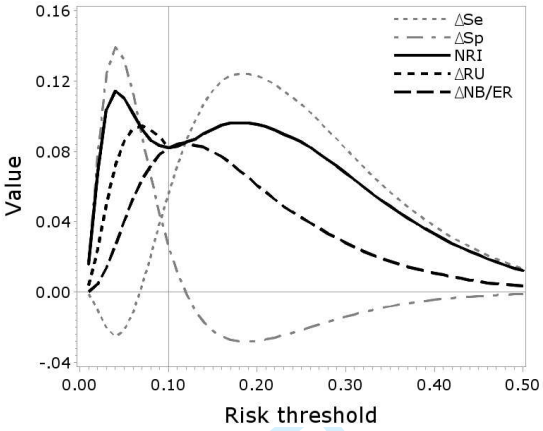
E (event rate 30%)



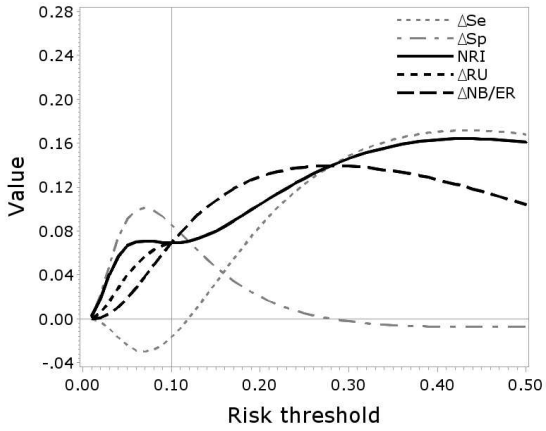
F (event rate 70%)



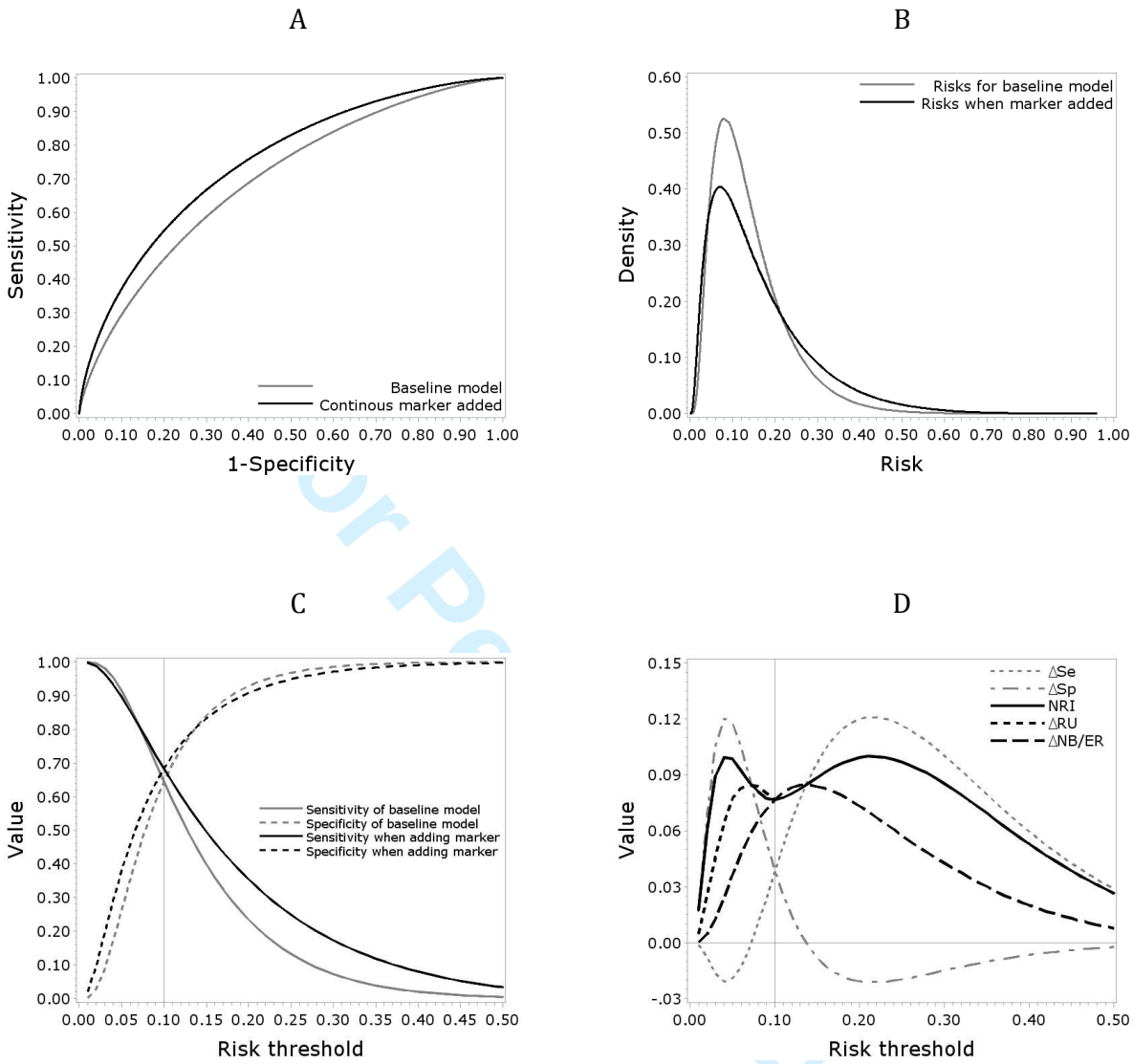
G (higher prevalence of marker)



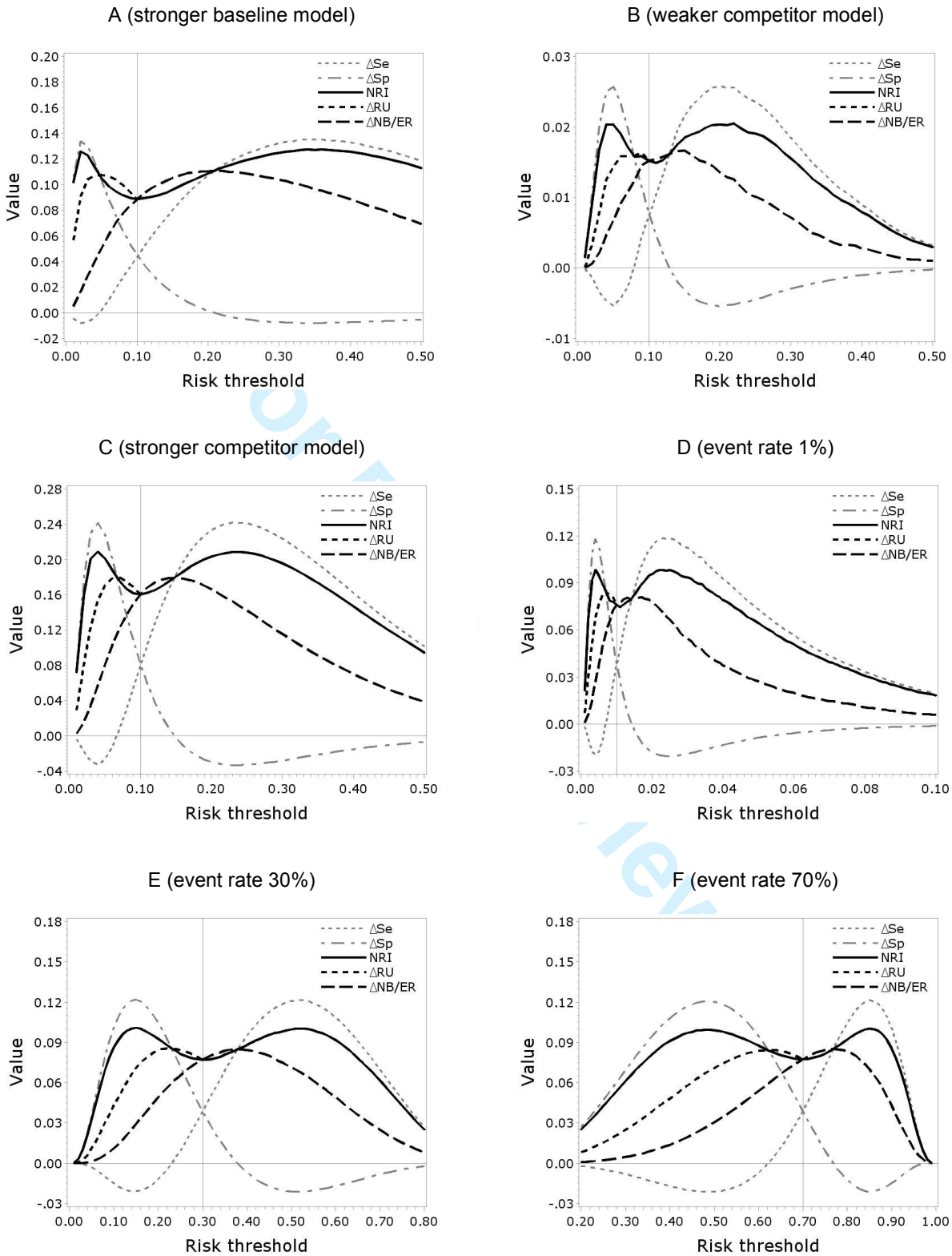
H (lower prevalence of marker)



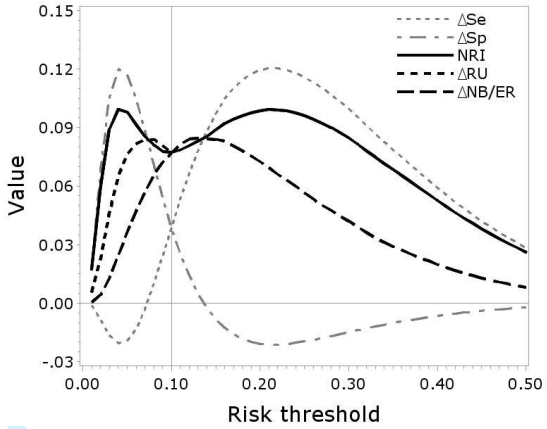
Web Figure 3.


A
Sensitivity
1-Specificity
Baseline model
Continous marker added
B
Density
Risk
Risks for baseline model
Risks when marker added
C
Value
Risk threshold
Sensitivity of baseline model
Specificity of baseline model
Sensitivity when adding marker
Specificity when adding marker
D
Value
Risk threshold
ΔSe
ΔSp
NRI
ΔRU
ΔNB/ER

Web Figure 4.

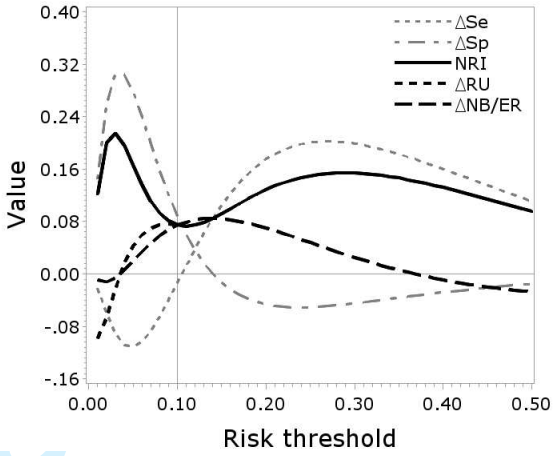

A (stronger baseline model)
B (weaker competitor model)
C (stronger competitor model)
D (event rate 1%)
E (event rate 30%)
F (event rate 70%)
ΔSe
ΔSp
NRI
ΔRU
ΔNB/ER
Value
Risk threshold

(G) higher correlation between competing models

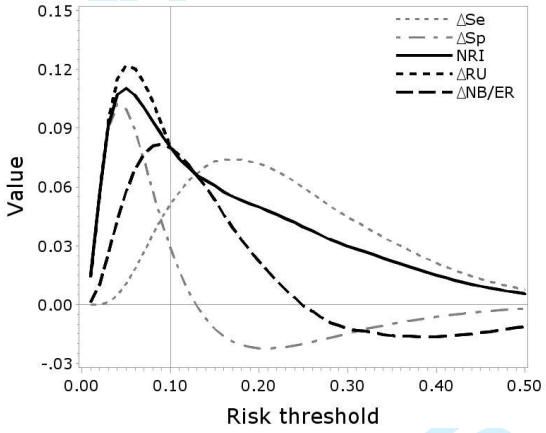


Web Figure 5.

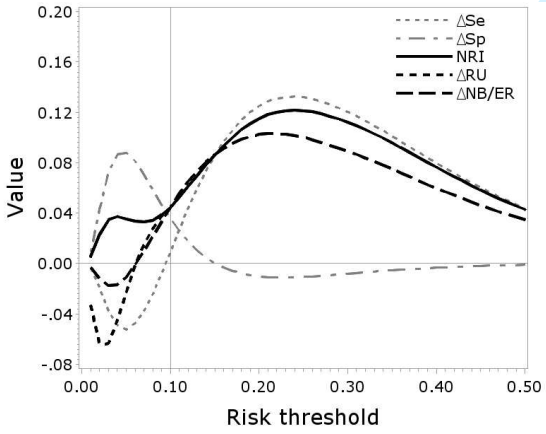
A (model with marker is overfit)



B (missed quadratic effect of marker)



C (missed interaction between model and marker)



Web Appendix 1: Settings for scenarios II (adding a binary marker to a prediction model) and III (comparing non-nested prediction models)

Adding a binary marker – The main difference here is that X_2 is now binary. This is obtained by dichotomizing a normally distributed variable. The AUCs of the baseline and extended models is set to the same values as in the main setting of scenario I. In the main setting the prevalence of the binary marker among non-events is set at 15%, whereas the prevalence among events is 41% to obtain the desired ΔAUC of 0.05. Again, X_1 and X_2 are uncorrelated and event rate is 10%. Variations considered are (Table A1): (a) the baseline model has an AUC of 0.8 while keeping ΔAUC at 0.05; (b) the added marker has a weak incremental value: prevalence among events 25% and ΔAUC is 0.01; (c) the added marker has a strong incremental value: prevalence among events 56% and ΔAUC is 0.1; (d) event rate is 1%; (e) event rate is 30%; (f) event rate is 70%; (g) marker prevalence among non-events is 50%, such that prevalence among events is 78% to retain the original ΔAUC of 0.05; (h) marker prevalence among non-events is 1% and among events is 19% to retain the original ΔAUC.

Comparison of non-nested models – In this situation X_1 and X_2 represent predictors from two competing and non-nested models. In the main setting, the competitor model based on X_2 has an AUC that is 0.05 higher than the AUC of the baseline model based on X_1 which is set at 0.7. As it is reasonable to assume that good competing models are correlated, we set the correlation between X_1 and X_2 to 0.5 among non-events as well as among events. The following variations are considered (Table A2): (a) the baseline model has an AUC of 0.8 while keeping ΔAUC at 0.05; (b) the competitor model has a small AUC advantage of 0.01; (c) the competitor model has a large AUC advantage of 0.1; (d) event rate is 1%; (e) event rate is 30%; (f) event rate is 70%; (g) correlation between X_1 and X_2 is 0.75.

Table A1. Settings for Assessing the Incremental Value of a Binary Marker

Event rate	Mean of X_1	Prevalence of marker among non-events and events	AUC Baseline model	AUC Extended model	ΔAUC
Main setting					
10%	0.742	15% and 41%	0.7	0.75	0.05
Variations					
Stronger baseline model					
10%	1.19	15% and 53%	0.8	0.85	0.05
Weaker or stronger added marker					
10%	0.742	15% and 25%	0.7	0.71	0.01
10%	0.742	15% and 56%	0.7	0. 8	0.1
Higher or lower event rate					
1%	0.742	15% and 41%	0.7	0.75	0.05
30%	0.742	15% and 41%	0.7	0.75	0.05
60%	0.742	15% and 41%	0.7	0.75	0.05
Higher or lower prevalence of marker					
10%	0.742	50% and 78%	0.7	0.75	0.05
10%	0.742	1% and 19%	0.7	0.75	0.05

AUC, area under the receiver operating characteristic curve; ΔAUC, difference in area under the receiver operating characteristic curve

Table A2. Settings for the Comparison of Non-nested Models

Event rate	Mean of X_1	Mean of X_2	Correlation X_1-X_2	AUC Baseline model	AUC Competitor model	ΔAUC
Main setting						
10%	0.742	0.954	0.5	0.7	0.75	0.05
Variations						
Stronger baseline model						
10%	1.19	1.466	0.5	0.8	0.85	0.05
Weaker or stronger competitor model						
10%	0.742	0.783	0.5	0.7	0.71	0.01
10%	0.742	1.19	0.5	0.7	0.71	0.1
Higher or lower event rate						
1%	0.742	0.954	0.5	0.7	0.75	0.05
30%	0.742	0.954	0.5	0.7	0.75	0.05
70%	0.742	0.954	0.5	0.7	0.75	0.05
Higher correlation between models						
10%	0.742	0.954	0.75	0.7	0.75	0.05

AUC, area under the receiver operating characteristic curve; ΔAUC, difference in area under the receiver operating characteristic curve

Web Appendix 2: Formulas for NRI, ΔNB/event rate, and ΔRU

The NRI, ΔNB/event rate and ΔRU can be written as a function of ΔSe and ΔSp [31]. If we denote the threshold by *T* and the event rate by *ER*, the formulas are as follows:

$$\mathrm{NRI} = \Delta\mathrm{Se} + \Delta\mathrm{Sp},$$

$$\Delta\mathrm{NB}/ER = \Delta\mathrm{Se} + \frac{odds(T)}{odds(ER)}\Delta\mathrm{Sp},$$

$$\frac{\Delta\mathrm{NB}}{odds(T)(1-ER)} = \frac{odds(ER)}{odds(T)}\Delta\mathrm{Se} + \Delta\mathrm{Sp},$$

$$\Delta\mathrm{RU} = \begin{cases} \dfrac{odds(ER)}{odds(T)}\Delta\mathrm{Se} + \Delta\mathrm{Sp}, \text{ if } T < ER \\ \Delta\mathrm{Se} + \dfrac{odds(T)}{odds(ER)}\Delta\mathrm{Sp}, \text{if } T \geq ER \end{cases}.$$

Web Appendix 3: beyond normality and the logistic regression model

In this appendix we extend the simulations to other algorithms than logistic regression, more specifically probit regression, Poisson regression, and support vector machines (SVM). In addition we investigate whether the results generalize to (1) a situation where the added marker is lognormal among events and non-events and (2) a situation, potentially less realistic, where the baseline model contains multiple binary markers and the added marker is binary as well. T An overview of the eleven settings addressed in this Appendix is provided in Table A3. We used simulated data of sample size 500,000, except for the SVM analyses where computational issues made us choose a sample size of 5,000. The results for the 'normal+normal' situation are shown in Figure A1, for the 'normal+lognormal' situation in Figure A2, and for the 'binary+binary' situation in Figure A3. These figures show that the main findings generalize to other prediction algorithms and to situations where the baseline and/or added markers are non-normal.

Table A3. Settings for Assessing the Incremental Value of a Binary Marker

Baseline (AUC)	Added marker (AUC)	Sample size	Event rate	Algorithm
Normal (0.70)	Normal (0.75)	500,000	0.1	Logistic regression
		500,000	0.1	Probit regression
		500,000	0.1	Poisson regression[1]
		5,000	0.1	Support Vector Machine[2]
Normal (0.70)	Lognormal (0.80)	500,000	0.1	Logistic regression
		500,000	0.1	Probit regression
		500,000	0.1	Poisson regression[1]
		5,000	0.1	Support Vector Machine[2]
4 binary variables[3] (0.63)	Binary[3] (0.70)	500,000	0.1	Logistic regression
		500,000	0.1	Probit regression
		500,000	0.1	Poisson regression[1]

Note: in each setting, the variables are uncorrelated.

[1] Following Zou [39].

[2] More specifically, a Bayesian least squares support vector machine with a linear kernel was used [40-42]. This analysis was done in Matlab version R2007a (www.mathworks.com) using the LSSVMlab toolbox.

[3] The four binary markers in the baseline model each have a prevalence of 15% among non-events and 25% among events. The added binary marker has a prevalence of 15% among non-events and 41% among events.

Figure A1. Assessment of improved classification in the 'normal+normal' setting. Using (A) logistic regression, (B) probit regression, (C) Poisson regression, or (D) support vector machines.

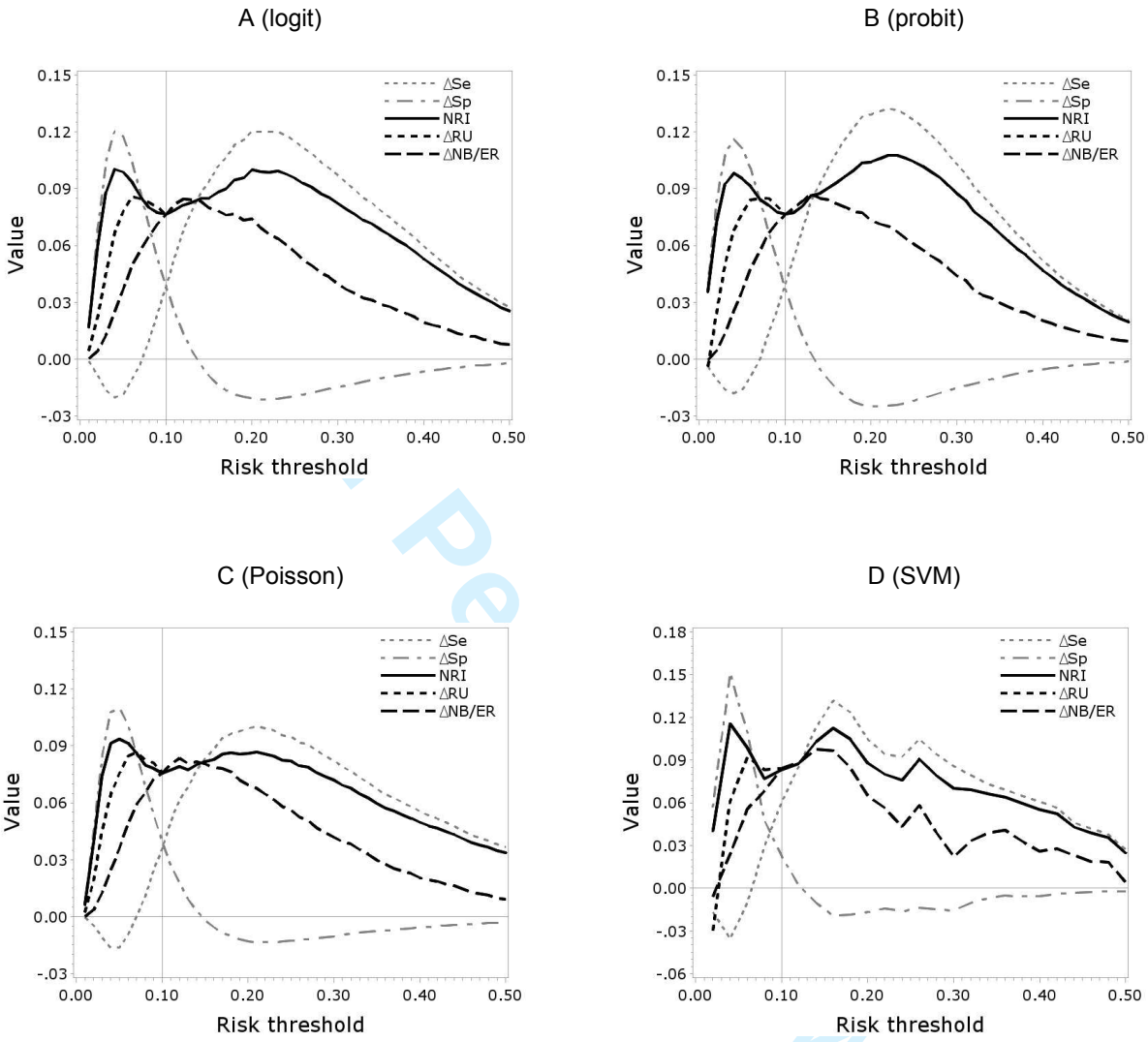


Figure A2. Assessment of improved classification in the 'normal+lognormal' setting. Using (A) logistic regression, (B) probit regression, (C) Poisson regression, or (D) support vector machines.

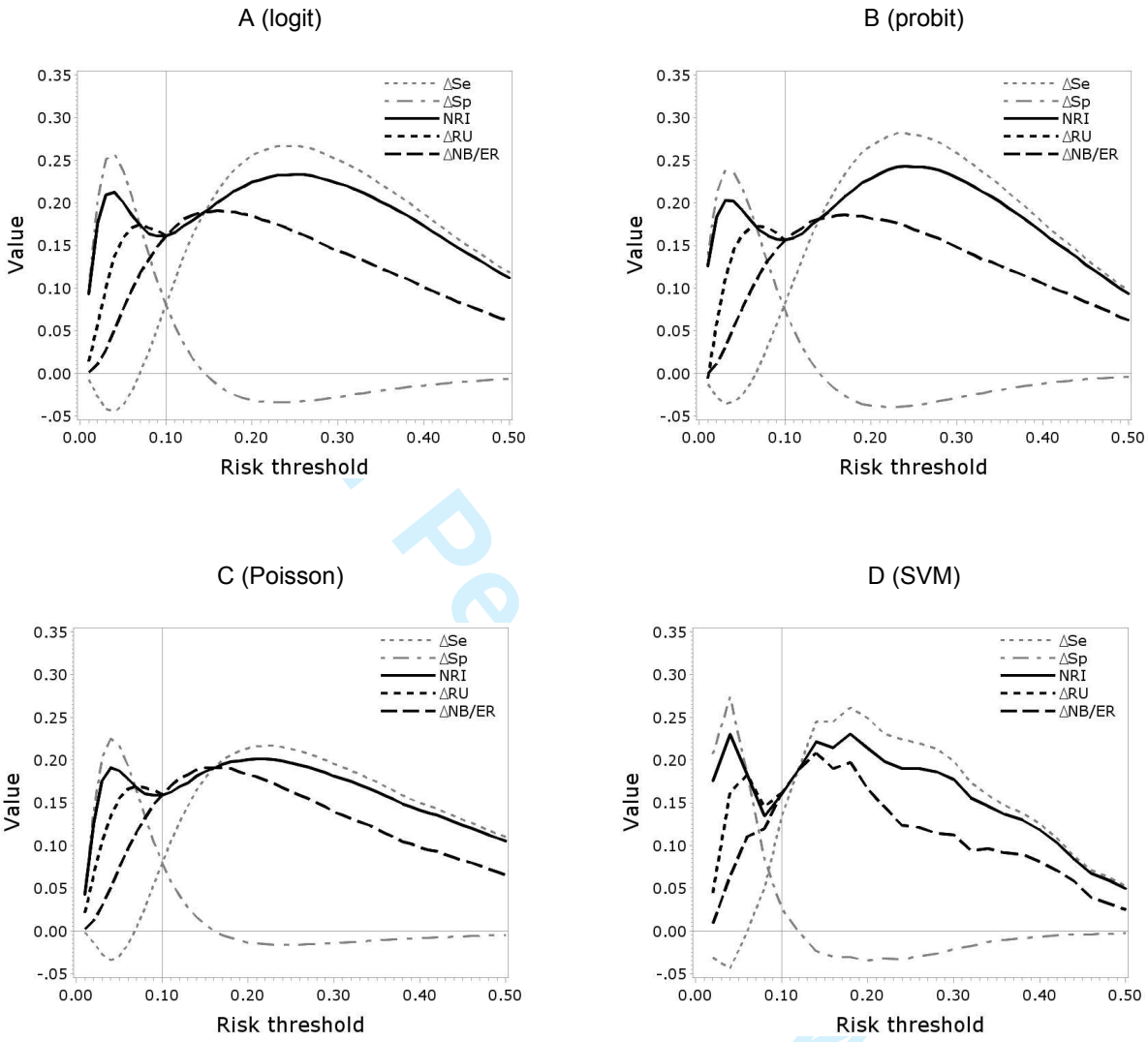


Figure A3. Assessment of improved classification in the 'binary+binary' setting. Using (A) logistic regression, (B) probit regression, or (C) Poisson regression.